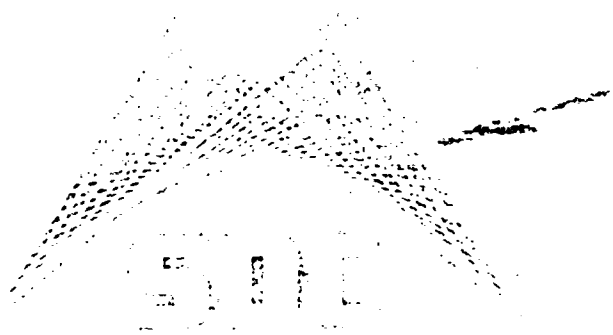

SP-1678

RAPID, a System for RETRIEVAL THROUGH AUTOMATED PUBLICATION AND INFORMATION DIGEST

Louise Schultz

1 June 1964

SP-1678/000/00

SP *a professional paper*

RAPID, a System for RETRIEVAL THROUGH AUTOMATED PUBLICATION AND INFORMATION DIGEST

by

Louise Schultz

1 June 1964

SYSTEM
DEVELOPMENT
CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA

RAPID, a System for RETRIEVAL THROUGH AUTOMATED PUBLICATION AND INFORMATION DIGEST

ABSTRACT

The exponentially increasing records of the technological advances of our times are overloading the systems available for communicating technical information. The two principal aspects of the communication need are improved dissemination and improved retrievability. The rise in numbers of technical documents has meant an increased publication backlog and longer delays in dissemination. It has also meant an increased "storage" backlog and longer delays in retrieval. More journals could be established to alleviate the publication backlog, but this solution would also increase the storage backlog.

This paper describes a conceptual model for an advanced information system that can improve dissemination without increasing storage problems. Referred to as RAPID (Retrieval through Automated Publication and Information Digest), the concept comprises a coordinated, semiautomated dissemination and retrieval system built around a special science newspaper. The RAPID concept is described in terms of its applicability to a technical and management community of broad subject base, large size, and geographic dispersion. A comprehensive, but preliminary, system diagram is provided in support of the generalized description.

Responsive to the recommendations of the President's advisors on scientific and technical information[4 and 9], the RAPID concept provides

> for reserving to the individual who produces information the responsibility for helping others use that information, by selecting appropriate indexing terms, with the guidance of system personnel and policies;
>
> for exploring new modes of information processing and retrieval;
>
> for supporting the Federal government in its efforts to increase compatibility among the various systems involved in the present information transfer network; and
>
> for augmenting scientific and technical information services.

Also responsive to Brunenkant's "pattern for information services progress"[3], the RAPID concept provides for timely dissemination--as a "market place for ideas," in terms appropriate for a wide range of readers; tends to centralize a large portion of technical information activities; and offers opportunities both for exploitation of a management dimension of technical information, and for research into natural language man-machine communication.

Implementation of the RAPID concept could effect a long range and beneficial reformation of technical communication and prove to be an invaluable tool to those responsible for managing the technical resources the system serves. RAPID could correct two of the inadequacies in the technical information transfer network that increasingly appear to threaten effective management of scientific resources--(1) the delay between occurrence of what may be called a technical event and the systematic availability of data about the occurrence to the general technical community, and (2) selective retrieval.

CAVEAT

This paper is cast, essentially, in a present-future tense rather than a conditional or propositional. The reader should not misinterpret this phraseology as implying that a RAPID system configuration exists or is imminent. Rather, the phrasing has been used in order to avoid the awkwardness of extensive use of auxiliary verbs such as "would" and "could."

Further, "RAPID" is not the name of a particular configuration of equipment and aggregation of procedures, but rather the label for a class of semi-automated information systems that achieves its mission by searching a data base it has accumulated principally from material it has also disseminated.

RAPID, a System for RETRIEVAL THROUGH AUTOMATED PUBLICATION AND INFORMATION DIGEST

Out of a combination of technical communication and technical management problems, out of the system approach to problem solution, and out of contemporary data processing technology has come a concept for coordinated, semiautomated technical information dissemination and retrieval. Called RAPID--for Retrieval through Automated Publication and Information Digest--the concept can be the basis for a system whose environment is characterized by

a user group large enough to justify mass communication techniques for dissemination of technical information;

a diversity in technical subject interests encompassed by the organization establishing and operating the system;

commitment by that organization to goals of effective technical communication and to exploration of innovative methods; and

a bond between sponsoring organization and the user group that facilitates data gathering and acceptance of system services.

A system implementing the RAPID concept has a dual mission, serving its users (1) by disseminating information rapidly and (2) by retrieving specific data from its store reliably. To achieve these missions, the system

gathers information pertinent to its service goals;

prepares this information for computer processing;

semiautomatically composes and publishes a dissemination vehicle such as a newspaper;

distributes this special-interest newspaper at least weekly;

organizes and digests the information published, to facilitate semiautomated retrieval of specific data or correlations of data;

maintains accurate characterizations of the group it serves; and

responds to individual technical and management queries--in its mature implementation--more rapidly and reliably than can any other retrieval technique.

The development of a RAPID system involves an analysis of system purposes, functions, and tasks. To achieve its purpose or mission, a system performs functions, each of which can be described in terms of a set of tasks. Tasks

are implemented by directed effort of humans, sometimes supported by mechanical and automatic tools and techniques, or else of machines. The objective of design is to specify the optimum combination of human and machine effort appropriate for each task. In a complex system, the performance of each task depends on the performance of one or more other tasks. Therefore, the designer must consider for each task each of the questions listed here, but not without regard to those that are lower on the list.

> How much of what kind of support does a human need to perform the judgmental or intellectual aspects of the specified task? For example, can the human perform required computations or logical correlations on the volume and kind of input data given, within the allowed period before task outputs are required for some other task?
>
> How much of what kind of support does a human need to perform the physical aspects of the specified task? For example, if the task inputs are visually accepted, is a device needed that improves image resolution, or that provides a means of varying the time or rate at which the human accepts inputs?
>
> Are support tools available or is their timely development feasible?
>
> How does the planned support for a particular task influence the performance of other tasks--how does the kind, volume, and rate at which the subject task produces data or material with the planned support, in differing from the case without this support, affect other tasks?
>
> What is the relative cost of planned support--in money, in time, and in the protection of human life, system components, and reliable system performance?
>
> What is the absolute cost of planned support--is it compatible with the commitments of the funding source?

In addition to serving information needs of both the manager and the technologist, the concept of Retrieval through Automated Publication and Information Digest serves a dual operation mission. Thus, interdependence among tasks is explicit in the RAPID concept.

To suggest the body of details involved in designing and implementing a system for Retrieval through Automated Publication and Information Digest, Figure 1, a conceptual model diagram, indicates the principal tasks or groups of tasks, organized into seven major functions: Input, Data Preparation, Current Awareness (newspaper production), Digest and Retrieval, Output, Self-Adaptation, and System Control.

RAPID system management and the user are represented at the left of the diagram. The blocks representing these elements cut the system diagram border to indicate that they are parts both of the system and its environment. The "user" may be providing management information, or he may be providing technical information. Anyone seeking information from a RAPID system for purposes other than managing the RAPID system, is a "user." The user, who generates much of the data accepted by the system, is constrained by system policies, and he is known to the system--two factors significant to successful system operation.

A NEWSPAPER FOR SCIENCE, TECHNOLOGY, AND MANAGEMENT

Some preliminary assumptions must be made about a given RAPID configuration, which are the basis for the description of this conceptual model. Output volume of the dissemination vehicle will be not less than 50,000 copies of a standard eight-column, (nominally) 15-inch by 23-inch newspaper, of eight pages (average), published at least weekly. Contents of each issue will comprise an average of 60 articles averaging 500 words each (including descriptors), headlines, and illustrations.

Portions of the system involved most directly in the dissemination mission and shown as individual blocks on the diagram are

the User (who provides data and interest descriptors),

the Input Data Control tasks,

the Editorial Processing,

the Semiautomated Composition tasks,

the Mechanical Production,

the Interest Index Maintenance tasks,

the Users' Interests-Issue Content Comparator,

the Contents Tags Generator and Printer, and

the Distribution of copies.

Indirectly involved are sources of policies that affect the user, retrieval capabilities that are exploited for editorial support, and recommendations arising from user reactions.

Style and content of the newspaper are influenced by the dual use of the data to be published. Using newspaper production techniques can be expected to minimize the cost of publication. Using mass communication techniques should make a RAPID dissemination vehicle more effective than are other technical communication media, for a wider audience. However, newspaper production and communication techniques must be modified (1) to permit use of the published text for semiautomated retrieval without additional preprocessing, and (2) to develop habits of use that will minimize discrepancies between what the system needs and what the source of system inputs provides. For example, indexing terms needed for retrieval processing are to be printed with each item, either as a list or within text. In the latter case, the appearance of the term would differ from the balance of the text, e.g., terms may be italicized or boldface. Thus, the reader is informed (nonredundantly) that the term is an entry for retrieval. To have occurred in text, the variation would have been coded on the tape controlling the photocomposing equipment. The same code may constitute the tag identifying a term to be used for semiautomated retrieval. A separate list of terms, possibly at the beginning of the text, though redundant, has the advantage that less total text need be scanned in generating the headline-bearing mailing labels (described later in this section). This consideration is not critical in a volume of text such as has been assumed, unless the Comparator cannot be fully automated. It would be significant were the interest indexes maintained on punch cards, for example.

In developing the indexing vocabulary, a RAPID system benefits from closed-system status: System policies, transmitted to system contributors who are part of the technical and administrative community, can require that data submitted for publication be accompanied by a set of terms useful for retrieval. As the system phases into operation, the set of allowable terms can be expected to change. System policy guides to contributors would reflect such changes.

The portion of the text contents contributed by the users, through their employers and in accordance with requirements, must be processed by Input Data Control, where acceptable data are routed to Editorial Processing. The Liaison task, shown as an output function, may contact users to clarify data that are not compatible with system requirements. The Input Data Control task may be performed manually or could be automated to provide that user contributions enter RAPID on data link and that data could be inspected, accepted or rejected, and routed by an appropriately programmed automatic data processing system.

To prepare accepted data for publication and use by the retrieval functions, Editorial Processing

edits;

rewrites;

gathers additional data--from the system file through the Query

Analysis and Format task, or from outside the system through the Liaison task;

prepares illustrations;

writes headlines;

makes up the page layouts, etc.

The portion of the text contents acquired entirely on the initiative of the system includes features on specific work and individuals; on legislation, national and international events, and business trends that affect the user group; and sponsoring agency's editorial and policy statements. On the average, this combination of noncontributed material should constitute about one third of each issue.

A RAPID system would prepare data for use both for publication and, without significant additional processing, for machine storage. Contemporary techniques of semiautomated composition, being developed in several metropolitan newspapers[2] and being investigated for use in technical communication, are appropriate for this task.

A RAPID publication vehicle is a current awareness tool, per se. Nevertheless, because each published item carries its set of descriptors, the usefulness of a given issue to the individual user can be augmented by providing--perhaps as part of the mailing label--a list of the headlines for items likely to interest that user. To provide this service, the system performs two or three tasks whose cost is justified by increased technical user service and increased management information. These tasks include gathering and maintaining sets of descriptors characterizing the interests of each user, comparing these sets against the descriptors in the issue, and generating and printing out the headline-bearing mailing labels. The service (1) appeals to the user as personalizing a mass communication medium; (2) provides system internal management with an indicator of user interests that facilitates improvements in coverage and internal data base organization; (3) permits the system to provide a channel between users with similar interests; and (4) constitutes a source of information for external management.

To generate the individual headline-bearing mailing labels, or contents tags, descripters profiles for a statistically significant portion of the user group must be established and maintained. Although intended to be more specific than a classification, the interest descriptors are not to be so specific that only one article in a year's collection of issues is brought to the user's attention. The goal should be to flag for the user sample an average of one or two items per issue. Interest profiles may be limited in number of entries. The Comparator processor may be programmed to look for no more than two hits (for this user and this issue) and to cycle descriptors that have hit for this issue to the bottom of the list (of this user's interests), etc.

In an eight-page publication, individual contents notices might be considered unnecessary--if the contents were comparatively homogeneous or if the publication were to be used for reference. But, on the assumption that the publication is intended primarily for rapid dissemination of a broad range of subject matter and constitutes an opportunity for stimulating exposure, backed up by a semiautomated retrieval system, the contents tag increases information transfer. Clearly, RAPID's "selective dissemination" need not assume that the user will read only what is brought to his attention. Thus, effectiveness of the service is to be measured by the accuracy with which the total "profiles" index reflects the interests of the total user group, rather than by the accuracy with which a particular item matches the stated interests of a particular subscriber.

To ensure the usefulness of the contents tags and, therefore, of any conclusions drawn from the Interest Index File, the User Satisfaction Analysis conducts surveys and the Liaison task disseminates policy statements.

CITATIONS, EXCERPTS, AND INFORMATION DIGEST

At the right of the RAPID system conceptual model diagram are shown the digest and retrieval functions. Except for the contents tag generation tasks and the mechanical production and distribution of the newspaper, the tasks directly involved in the dissemination mission are also directly involved in the retrieval mission. Mechanical production of the newspaper is indirectly involved inasmuch as searches of the publication, on microfilm, are to be provided.

The newspaper medium offers speed of communication but, in general, does not offer systematic retrievability. To communicate effectively, the newspaper relies upon redundancy. Each item contains enough background data to be independent of previous items on the same subject. RAPID, designed to retrieve from published material, must be designed to cope with redundant text.

The system could be equipped with a unified storage capability retaining either the total text--and a storage organization method and search speed appropriate to retrieving from such a store--or portions selected by a "redundancy filter." In either case, every request is searched through the single store. However, we cannot expect all requests to require equal specificity or extent in the data they elicit from the store.

Therefore, the system is to be designed to match the expenditure on storage maintenance and search with the requirements of the query. Manual systems exhibit this characteristic; however, capabilities for storage and search speed in automated systems have masked the obvious advantages of storage hierarchies.

A hierarchy of storage is anticipated for RAPID, comprising a microfile of all published material, a "morgue," a detailed index or a concordance, an array of special indexes, and a (set of) catalog(s).* Except for the microfile and morgue, all stores or files are to be prepared and maintained in machine form. Semiautomated techniques are available for implementing these tasks. Descriptors in machine form for each issue can be accumulated into special indexes and used to generate catalogs. Searches of these stores provide citations to issues containing material responsive to a particular query. Text, also provided to the retrieval functions in machine form, can be processed by available computer programs to provide a concordance or a detailed index.

A concordance or a detailed index offers a capability for increasingly sophisticated search correlations both for editorial support of the system and for management and technical information synthesis. Searching a deep index elicits only citations; information synthesis can be accomplished only after reference to the contents of the material cited. Searching a cooccurrence index[7] or a concordance elicits some information.

With a deep index or a concordance, the system can provide a test bed for research into automated processing of natural language text. In addition to fundamental logical processing in matching individual terms, the system could be designed to generate and store syntactic tags for list entries[1]. Similar tags associated with constituents of the query statement increase the specificity with which the data base is searched. Although present research of this nature is promising, it has revealed the limitations of syntax as a basis for language analysis and synthesis. Hence, greater dependence on human comprehension of natural language (than on machine processing) is recommended for a RAPID system.

To provide an internal system capability for information analysis and synthesis [illegible] for research in automating these functions, a "morgue" would be included. Although the morgue would be maintained in machine form, it will be generated manually until sophisticated automatic abstracting techniques become available. In initial implementation, the morgue tasks are similar to specialized literature analysis and can be performed adequately only by highly skilled professionals whose effectiveness is augmented by the support offered by the semiautomated files.

Conventionally, literature analysis is "mission-oriented." Search sifting and analysis, as well as synthesis, are performed in response to a specific need.

---

* The parenthetic notes on the conceptual model diagram about the kind of storage medium for each of the various files are tentative suggestions. Alternates include putting all machine files on a single type of medium (tape, drum, disc, magnetic [illegible] decks, etc.).

Because the RAPID concept centers around a closed system, systematic development of an information-rich morgue is appropriate (1) to enhance the system's capability for verifying and amplifying input data, (2) for initiating significant data gathering, (3) for reducing time on individual queries, and (4) for providing the opportunity for research in automatic literature processing.

Because RAPID gathers and disseminates management as well as timely technical information, the retrieval functions can respond to questions characterized as "science intelligence" queries. Further, the "trail" left by a project can be examined even if the project does not culminate in success and publication of a "final report" or journal article. (Comprehensive publication of the details of an unproductive approach is neither practical--in view of the prospects of documentation inundation--nor professionally desirable. However, private communication to assist a peer and to prevent the waste of critical technical manpower and time is both practical and professionally desirable.)

In providing retrieval services, a RAPID system must be selective with respect to the identity of the questioner for whom the system could answer various kinds of questions and the depth of search service to be offered in each case. The Query Analysis and Format task may be the single element of the system having access to all system data contents (although it is not an output task and may be checked within the system by the Response Transmittal task). Query Analysis and Format selects the search strategy--in effect, setting the switches shown as the Query Router. The File selector switch may activate a search of (clockwise around the contacts)

the microfile, when the citation is included in the query;

the morgue, when information, rather than citations, is sought;

the concordance or detailed index, when unforeseen correlations are needed;

the special index selected by the Index Selector switch, or

the catalog.

To aid Query Analysis, the catalog search system includes a visual display device on which the portion of the catalog being referenced can be examined (and, if necessary, reproduced). A similar viewing station is suggested for the Text Selection Processor task. Having determined which indexes to examine, from the catalog, the search continues through the selected index, to generate a list of citations responsive to the descriptor set. Rather than depending upon promised development of an automatic search program as capable as a competent special librarian, a RAPID configuration should depend upon human post-editing. Thus, if a query is judged to justify the service, the

citation list generated by the search can be routed to a post-editor. Further, depending upon the decision of the query analyst, the citations (either with or without post-editing) can be used as the basis for a "fact" search. (If a deep index, rather than a concordance, is provided in the system, the selection is made at the index selector level.) Here the concepts of a hierarchy among queries and provision of a comparable hierarchy in search effort is important.

PLAN FOR SYSTEM DEVELOPMENT AND MODIFICATION

Contemporary research into automating information services is rich in the promise of designing and programming computer-based systems that will provide direct and immediate communication between the individual who needs information and a data base comprehensive enough to supply that information. The individual with the problem today prefers a solution less rich in promise but with a more readily estimated probability of performance. The primary policy guiding development and implementation of a RAPID configuration should be (1) determine the functions the system is expected to perform; then (2) design the tasks performing that function to be implementable immediately (within the phase-in considerations outlined) but provide the opportunity for design modification responsive to state-of-the-art advances, to changes in the sponsor's commitment, and to changes in specification of the function to which the task contributes.

The conceptual model includes, for example, a task called "Text Selection Processor." This does not imply that an automatic data processing program has been designed or is expected to be available for installation in a given configuration of the RAPID system. Rather, inclusion of that task block implies that the task of selecting text for retention is to be provided, and will be accomplished--initially, by a skilled literature analyst supported by a semi-automated file, perhaps, ultimately, by a semiautomated processor guided by a skilled literature analyst, both exploiting an automated file.

The next few paragraphs summarize an implementation logic appropriate to a development program of approximately two years. Based on design requirements of the retrieval functions, dissemination functions are implemented first, using a comparatively arbitrary set of descriptors. While the operational specifications for the newspaper subsystem are being prepared, the newspaper staff begins work on policy guides, to be provided to data sources; sets up the system for gathering data and processing inputs; and establishes the identity and mailing addresses of potential recipients of the newspaper. The programs for implementing automated aspects of the function are started. Equipment needs are determined and filled.

To develop a data base rich enough to justify searching will require publication and dissemination for a period of perhaps a year prior to acceptance of queries. Searches are performed in response to requests from the system's editorial

personnel and management during the early periods of operation, on the understanding that such exercises aid in developing specifications for ultimate system performance, and that inadequacies in the response are not to be interpreted as permanent or inherent performance characteristics.

After three or four months of publication, the newspaper announces its anticipated "personalized" service and provides the subscriber with a coupon for specifying descriptors he believes characterize his interests. He will not be limited to those that have appeared in the newspaper. The Interest Index Maintenance task enters a development phase lasting three to six months, in which the user profiles and newspaper contents are analyzed. Although all subscribers should be solicited for profiles, initial task implementation will be based on analysis of a significant sample for which the Comparator Processor and Tag Generator will produce mailing labels. Close cooperation by this sample of users will provide the details for routine operation of the service.

Text published is available in machine form. During the first few months of publication, a comparatively simple descriptor indexing system is operated, principally as an internal editorial support and system development tool. Development of the data processing programs and implementation of system equipment for query analysis and response are completed during the initial months of publication. Development of vocabulary and search techniques parallel the development of the contents tags service. Routine query service, including post-editing of citation lists and transmittal of lists or of reproductions of cited items, can be offered early in the second year of publication.

To develop the morgue and, ultimately, to synthesize published data for enhancing system performance and in response to queries, the Text Selection task is to be established early in the publication period. Techniques for selecting text fragments to be retained and for facilitating selection, by reference to morgue contents, are developed and increasingly automated during the first two or three years of publication. As the significance of this data base grows, the capability for exploiting it will grow. The alternative--searching an existing nonselectively accumulated data base in response to specific queries (the present literature search situation)--has major weaknesses and should not be considered for a coordinated information system[5].

By the end of the first six months of publication, the satisfaction of users with the medium should be measured. To set up an ongoing user satisfaction survey program, the User Satisfaction task will be implemented. The Answer Evaluator task need not be set up until after the first year of publication. At that time, the System Modification Center is to be established for coordinating operations data, user data (from interest indexes), user satisfaction data, and answer evaluation. Although analysis of the returns from user satisfaction surveys involves automatic data processing support, the Self-Adaptive functions are performed primarily by a team of researchers.

(Page 14 blank)

CONCLUSION

Information system research is and must remain characterized by specialization --by the mission to reveal or establish principles by dissecting ever smaller parts of the process of communicating. Applying such research largely echoes variations in progress of independent projects. The result is continuation of the policy of improving components rather than designing systems. A system approach to communication, in a milieu of increasingly capable techniques and equipment, is taken in the RAPID concept. To accomplish a system breakthrough[6] now requires the initiative of an agency sensitive to the opportunities for improved communication--if not suffering from poor communication --and to federal responsibilities for communication of technological and scientific information. We need not "wait for Mr. Know-it-all"[8]. Today's information system technology and the return to the user of responsibility for exposure, if not specialized search, promise significant improvement in a presently tangled and discontinuous information transfer network.

REFERENCES

1. Bobrow, D. G., Syntactic analysis of English by computer--a survey, AFIPS Conference Proceedings, v. 24, 1963, pp. 365-87.

2. Booth, O., A new L. A. story, Quill, May 1963, pp. 8-12.

3. Brumenkant, E. J., Information services--a pattern for progress, presented at the 10th Pacific Science Congress, Honolulu, 25 August 1961.

4. Crawford, J. H., ed., Scientific and technological communication in the government. Washington: April 1962, AD299545.

5. Doyle, L. B., Expanding the editing function in language data processing, SP-1266. Santa Monica, Calif.: System Development Corp., 10 July 1963.

6. Doyle, L. B., How to plot a breakthrough, SP-1492. Santa Monica, Calif.: System Development Corp., 12 December 1963, pp. 18-19.

7. Schultz, L., Retrieving untitled documents, SP-1637. Santa Monica, Calif.: System Development Corp., 5 May 1964.

8. Way, K., N. B. Gove, and R. van Lieshout, Waiting for Mr. Know-it-all, Physics Today, February 1962.

9. Weinberg, A., ed., Science, government, and information. Washington: GPO, 10 January 1963.

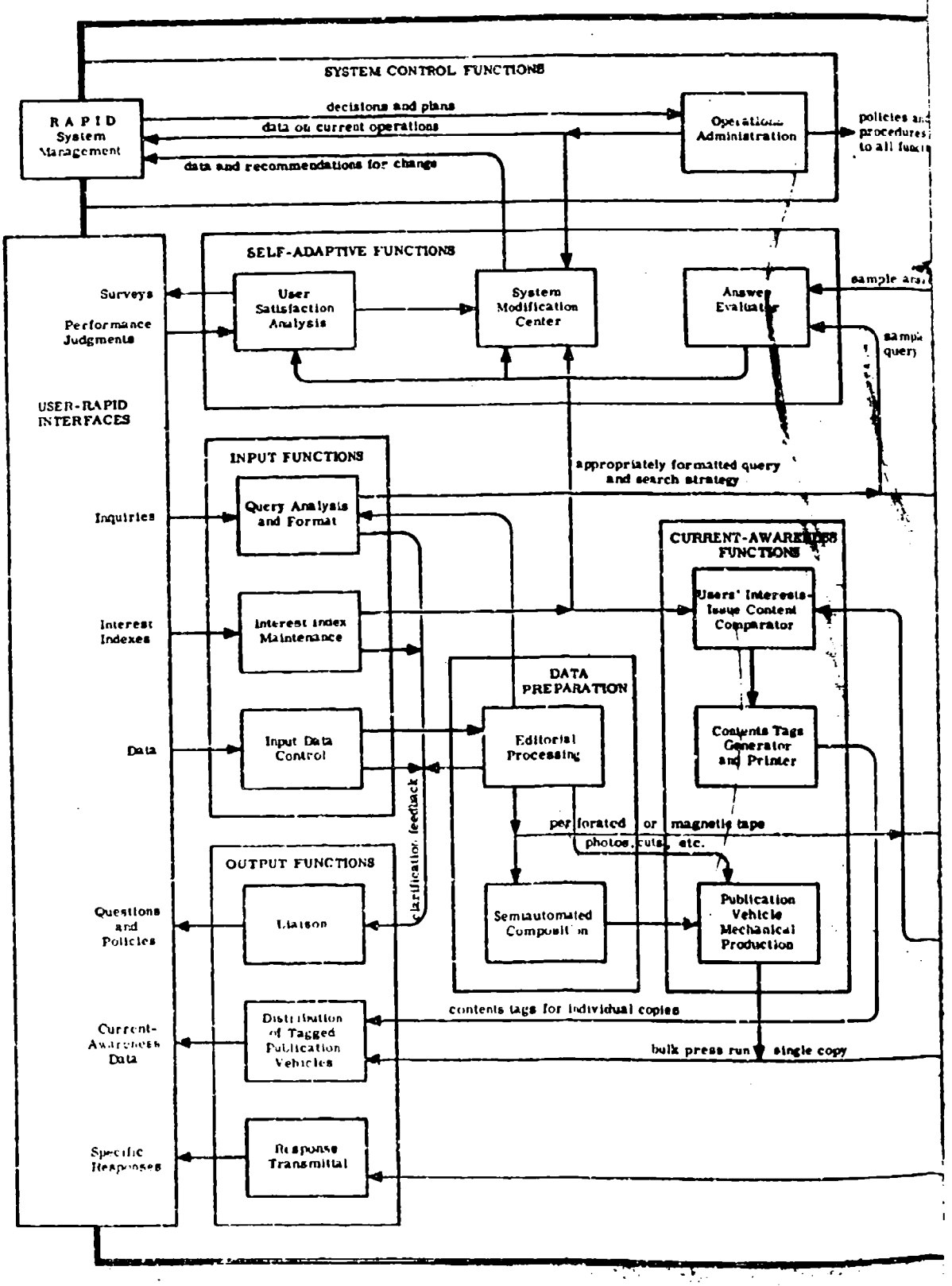
SYSTEM CONTROL FUNCTIONS
RAPID System Management
decisions and plans
data on current operations
data and recommendations for change
Operations Administration
policies and procedures to all functions
SELF-ADAPTIVE FUNCTIONS
Surveys
Performance Judgments
User Satisfaction Analysis
System Modification Center
Answer Evaluator
sample area
sample query
USER-RAPID INTERFACES
INPUT FUNCTIONS
appropriately formatted query and search strategy
Inquiries
Query Analysis and Format
CURRENT-AWARENESS FUNCTIONS
Users' Interests-Issue Content Comparator
Interest Indexes
Interest Index Maintenance
DATA PREPARATION
Data
Input Data Control
Editorial Processing
Contents Tags Generator and Printer
perforated or magnetic tape
photos, cuts, etc.
clarification feedback
OUTPUT FUNCTIONS
Questions and Policies
Liaison
Semiautomated Composition
Publication Vehicle Mechanical Production
contents tags for individual copies
Current-Awareness Data
Distribution of Tagged Publication Vehicles
bulk press run
single copy
Specific Responses
Response Transmittal

FIGURE 1

A PRELIMINARY CONCEPTUAL MODEL OF

AN SDC SYSTEM FOR

RETRIEVAL THROUGH AUTOMATED PUBLICATION AND INFORMATION DIGEST

RAPID